Recursive PAC-Bayes

Yevgeny Seldin

Post-Bayes Seminar Series

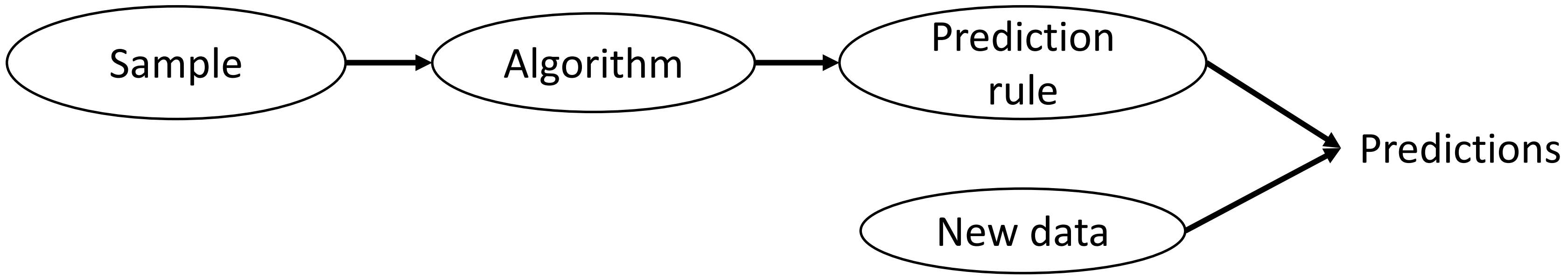
7 October 2025

Outline

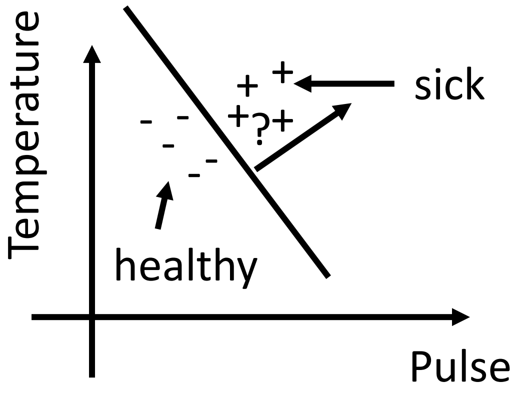
- Supervised learning – general background on generalization guarantees
- Occam’s razor – “the little brother of PAC-Bayes” [skipped]
- PAC-Bayesian analysis (including distinctions with Bayesian learning)
- Recursive PAC-Bayes – sequential prior updates
- Weighted majority votes (if we reach it…)

Supervised Learning

- Protocol

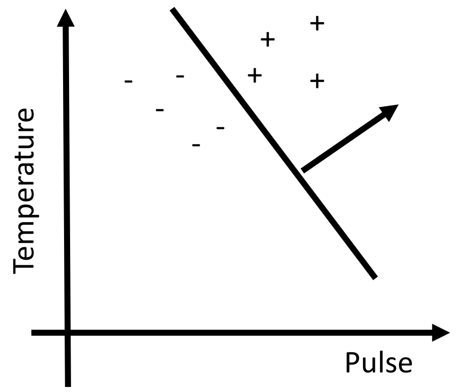


- Example



Supervised Learning

- Notations
  - $\mathcal{X}$ – sample space (e.g., $\mathcal{X} = \mathbb{R}^d$)
  - $\mathcal{Y}$ – label space (e.g., Classification: $\mathcal{Y} = \{\pm 1\}$; Regression: $\mathcal{Y} = \mathbb{R}$)
  - $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ – training sample (where $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$)
  - $h: \mathcal{X} \to \mathcal{Y}$ – a prediction rule / hypothesis
  - $\mathcal{H}$ - a set of prediction rules / a hypothesis set



Evaluation

- $\ell(y', y)$ – loss/error/risk function
  Loss for predicting y' when the reality is y

- Examples:
  - Zero-one loss
    $$\ell(y', y) = \mathbb{I}(y' \neq y) = \begin{cases} 0, & if\ y' = y \\ 1, & if\ y' \neq y \end{cases}$$
  - Squared loss
    $\ell(y', y) = (y' - y)^2$
  - Absolute loss
    $\ell(y', y) = |\mathrm{y}' - \mathrm{y}|$

- **The loss function determines the cost of different mistakes!!!**

- Example: Fire alarm

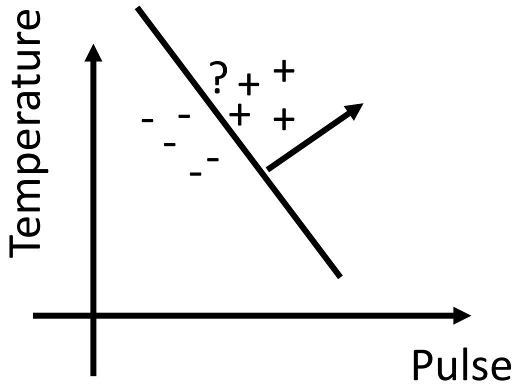
Depends on the house

y' \ y	no fire	fire
no fire	0	5.000.000
fire	2.000	0

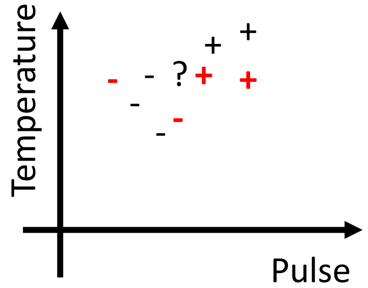
“constant”

The quantity of interest – Expected Loss

- Expected loss/error/risk
  - $L(h) = \mathbb{E}_{(X,Y)\sim p(x,y)}[\ell(h(X), Y)]$
- Assumption
  - (X, Y) are sampled from a fixed (unknown) distribution $p(X, Y)$
- Challenge: $p(X, Y)$ is unknown, and so is $L(h)$
- What can we say about $L(h)$?
  - Use empirical loss $\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$ as an estimate
- **Key question: how close is $\hat{L}(h, S)$ to $L(h)$?**



Validation



$\boxed{(X,Y)}$

$$S = \overbrace{\underbrace{S_{train}}_{\downarrow\, h}}^{n-m}\ \overbrace{\underbrace{S_{val}}_{\downarrow}}^{m}, \quad |S| = n$$

$$\hat{L}(h, S_{val}) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(X_i), Y_i)$$

- **Assumptions**
  - $\{(\boldsymbol{X_1}, \boldsymbol{Y_1}), \dots, (\boldsymbol{X_m}, \boldsymbol{Y_m})\}$ **are independent identically distributed (i.i.d.)**
  - **And come from the same distribution as new samples** $(\boldsymbol{X}, \boldsymbol{Y})$

- $\hat{L}(h, S_{val})$ is an **unbiased** estimate of $L(h)$
  - $\mathbb{E}\left[\hat{L}(h, S_{val})\right] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \ell(h(X_i), Y_i)\right] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[\ell(h(X_i), Y_i)] = L(h)$
  - From the perspective of h the samples in S_{val} are indistinguishable from new samples (X, Y)

What can be said about $L(h)$ based on $\hat{L}(h, S_{val})$?

- $\hat{L}(h, S_{val})$ is an unbiased estimate of $L(h)$

- But consider $m = 1$ and the zero-one loss:
  - $\hat{L}(h, S_{val}) \in \{0,1\}$ – never close to $L(h)$!

- Being unbiased is neither sufficient, nor necessary

- We need concentration!

Formalization

- $Z_i = \ell(h(X_i), Y_i)$

- If $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent identically distributed (i.i.d.), then $Z_1, \dots, Z_n$ are also i.i.d.

- $L(h) = \mathbb{E}[\ell(h(X), Y)] = \mathbb{E}[Z_1] = \mu$

- $\hat{L}(h, S_{val}) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) = \frac{1}{n} \sum_{i=1}^{n} Z_i = \hat{\mu}_n$

- How far can $\hat{\mu}_n$ be from μ?

Frequentist vs. Bayesian paradigms

- Bayesian paradigm
  - Parameters of data-generating process are sampled from an unknown distribution
  - Bayesian learning starts with a prior distribution $\pi(\theta)$ on the parameters and, given evidence S, applies the Bayes rule
    - $\rho(\theta|S) = \frac{\pi(\theta)\mathbb{P}(S|\theta)}{\mathbb{P}(S)}$
  - The probabilities are over observations and parameters (both are random variables)
  - The loss function is not part of the basic framework!

- Frequentist paradigm
  - The parameters are unknown, but fixed
  - Frequentists bound the probability that some [loss] function of the observations $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$ deviates significantly from its expectation $L(h) = \mu = \mathbb{E}[\hat{\mu}_n]$
  - The random variable is $\hat{\mu}_n$, but not μ; and the probability is over $\hat{\mu}_n$, but not μ

Frequentist vs. Bayesian paradigms

PAC-Bayesian analysis takes the frequentist path

PAC = Probably Approximately Correct

- Frequentist paradigm
  - The parameters are unknown, but fixed
  - Frequentists bound the probability that some [loss] function of the observations $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$ deviates significantly from its expectation $L(h) = \mu = \mathbb{E}[\hat{\mu}_n]$
  - The random variable is $\hat{\mu}_n$, but not μ; and the probability is over $\hat{\mu}_n$, but not μ

Concentration of measure – Hoeffding's inequality

- Theorem (Hoeffding's inequality):
  Let $Z_1, \dots, Z_n$ be i.i.d., $Z_i \in [0,1]$, then for any $\varepsilon > 0$:
  $$\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}$$
  Equivalently, for any $\delta \in (0,1]$: $\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] \geq \frac{1}{n}\sum_{i=1}^{n} Z_i + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leq \delta$
- Corollary: Assume that ℓ is bounded in the $[0,1]$ interval. Assume that we have **a single** prediction rule h that is **independent** of S. Then for any $\delta \in (0,1]$:
  $$\mathbb{P}\left(L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leq \delta$$
  Equivalently, with probability at least $1-\delta$: $L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}$.

  The probability is over $\hat{L}(h,S)$, but not $L(h)$!!!

Learning by Selection

$$S = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

(X, Y)

$\mathcal{H}$

$$h_1 \rightarrow \hat{L}(h_1, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_1(X_i), Y_i)$$

$$h_2 \rightarrow \hat{L}(h_2, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_2(X_i), Y_i)$$

$$h_3 \rightarrow \hat{L}(h_3, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_3(X_i), Y_i)$$

$$\vdots$$

$$\hat{h}_S^* = \arg\min_{\mathrm{h} \in \mathcal{H}} \hat{L}(h, S)$$

$$L(\hat{h}_S^*) = \mathbb{E}[\ell(\hat{h}_S^*(X), Y)]$$

$$\boxed{\mathbb{E}_S[\hat{L}(\hat{h}_S^*, S)] \neq \mathbb{E}_S[L(\hat{h}_S^*)]}$$

Hoeffding does not apply to $L(\hat{h}_S^*) - \hat{L}(\hat{h}_S^*, S)$!

Hoeffding: $\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}$

Selection from finite $\mathcal{H}$ ($|\mathcal{H}| = M$)

$L(\hat{h}_S^*) \neq \mathbb{E}_S[\hat{L}(\hat{h}_S^*, S)]$

We cannot apply Hoeffding!

We break the dependence

$$\mathbb{P}\left(L(\hat{h}_S^*) \geq \hat{L}(\hat{h}_S^*, S) + \varepsilon\right)$$

$$\leq \mathbb{P}\left(\exists h \in \mathcal{H}\colon\ L(h) \geq \hat{L}(h, S) + \varepsilon\right)$$

(Union bound) $\quad \leq \sum_{h\in\mathcal{H}} \mathbb{P}\left(L(h) \geq \hat{L}(h,S) + \varepsilon\right)$

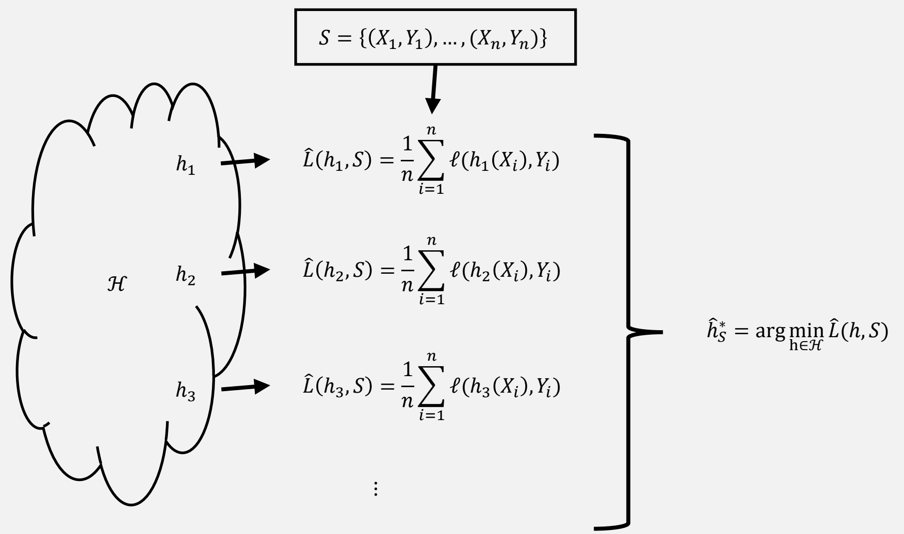
(Hoeffding) $\quad \leq \sum_{h\in\mathcal{H}} e^{-2n\varepsilon^2}$

$$= \underbrace{M}_{\text{Selection}} \times \underbrace{e^{-2n\varepsilon^2}}_{\text{Concentration}}$$

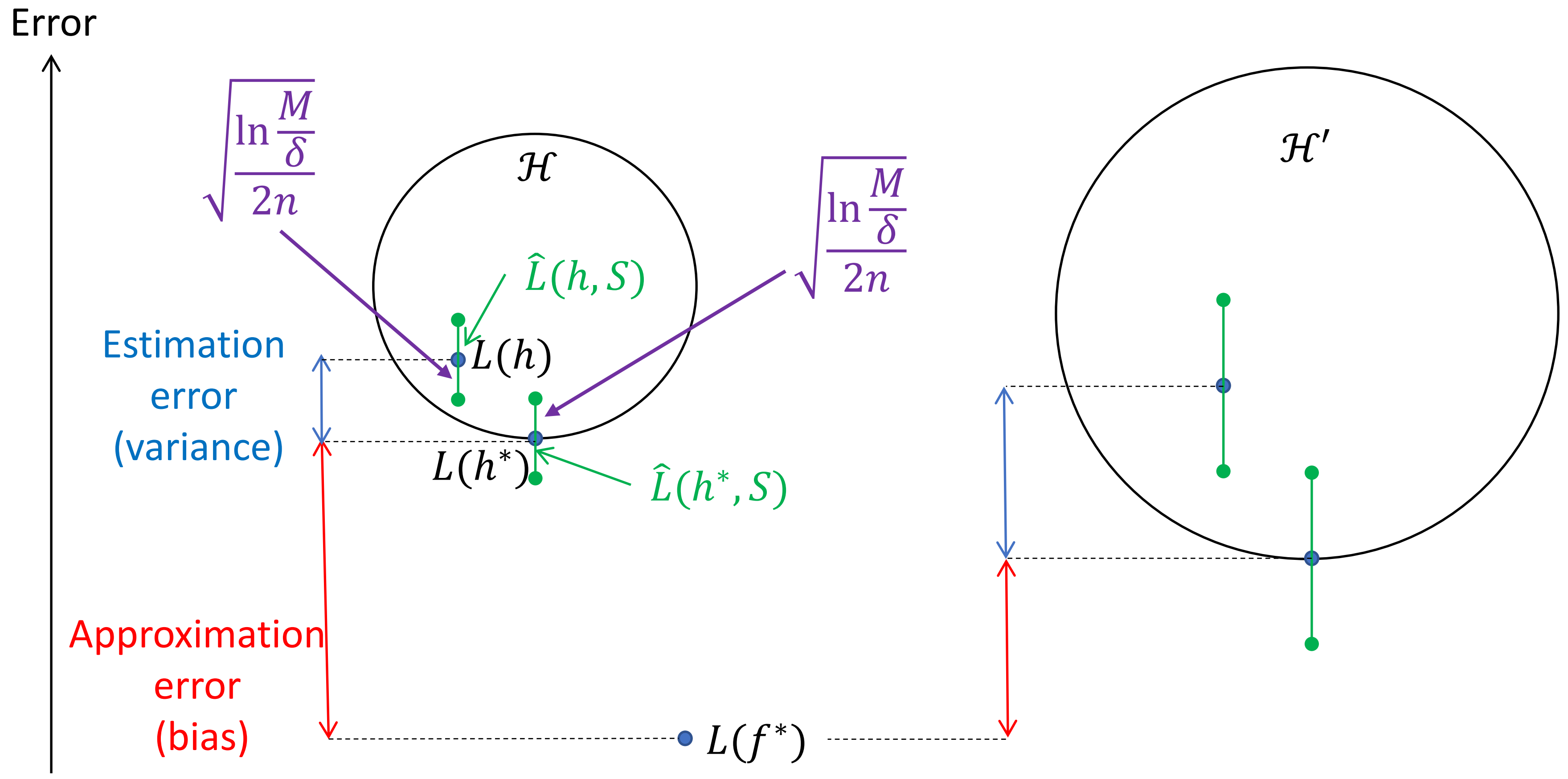
$$= \delta$$

Solving for ε gives $\varepsilon = \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}$

Theorem: $\mathbb{P}\left(L(\hat{h}_S^*) \geq \hat{L}(\hat{h}_S^*, S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right) \leq \delta$



Approximation-Estimation (bias-variance) trade-off



Estimation error $L(h) - L(h^*)$ can be up to $2\sqrt{\frac{\ln\frac{M}{\delta}}{2n}}$

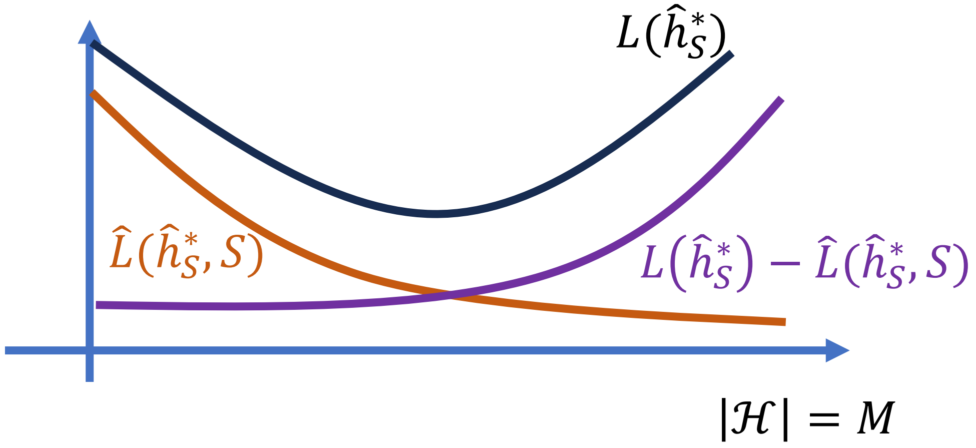
The more is not necessarily the better

Selection from a small $\mathcal{H}$ Selection from a large $\mathcal{H}$

Mid-summary

- $\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \underbrace{\sqrt{\frac{\ln\frac{M}{\delta}}{2n}}}_{\varepsilon}\right) \leq \underbrace{M}_{\text{Selection}} \times \underbrace{\frac{\delta}{M}}_{\substack{\text{Concentration} \\ e^{-2n\varepsilon^2}}} = \delta$

- For $M \ll e^n$ we have $L(h)$ under control
  - Concentration is stronger than selection



- Approximation-Estimation trade-off:
  - How to hit the "sweet spot"?
  - We have to pick $\mathcal{H}$ before we start working with the data!
- Can we let the data speak for itself?

Outline

- Supervised learning – general background on generalization guarantees
- **Occam's razor – "the little brother of PAC-Bayes" [skipped]**
- PAC-Bayesian analysis (including distinctions with Bayesian learning)
- Recursive PAC-Bayes – sequential prior updates
- Weighted majority votes (if we reach it…)

Occam's razor – "The little brother of PAC-Bayes"

- Occam's razor – adaptive selection from countable $\mathcal{H}$
  - A gentle introduction to "priors" in the frequentist framework.

- Check
  - Yevgeny Seldin. Machine Learning. The science of selection under uncertainty. https://arxiv.org/pdf/2509.21547, 2025.

Hoeffding: $\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leq \delta$

Occam's razor – Generalization bound for countable $\mathcal{H}$

- Theorem (Occam's razor): Let $\pi(h)$ be nonnegative and **independent of S** and satisfy $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$. Then:

$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta.$$

The bound for a finite $\mathcal{H}$ is a special case

$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right) \leq \delta$$

$$\pi(h) = \frac{1}{M}$$

- Proof:

$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}\right)$$

(Union bound) $\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left(L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}\right)$

(Hoeffding, π is independent of S!) $\leq \sum_{h\in\mathcal{H}} \pi(h)\delta$

($\sum_{h\in\mathcal{H}} \pi(h) \leq 1$) $\leq \delta$

Occam's razor selection

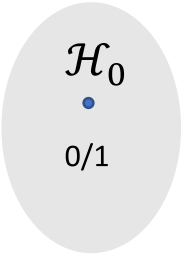
$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta$$

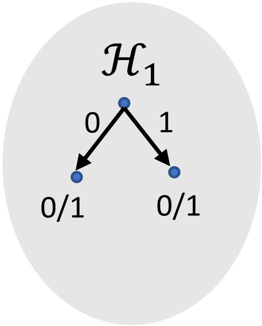
$$\mathbb{P}\left(\forall h \in \mathcal{H} : L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}\right) \geq 1 - \delta$$

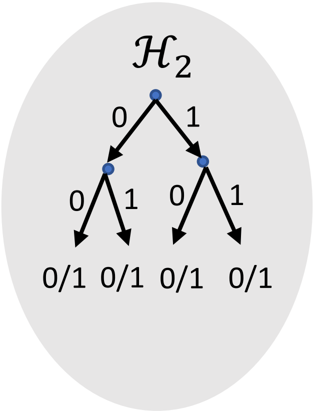
$$\hat{h}_S^* = \arg\min_{\mathrm{h}} \quad \underbrace{\hat{L}(h, S)}_{\substack{\text{Empirical} \\ \text{Performance}}} \quad + \underbrace{\sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}}_{\text{Complexity}}$$

With probability at least $1 - \delta$: $L\left(\hat{h}_S^*\right) \leq \hat{L}\left(\hat{h}_S^*, S\right) + \sqrt{\frac{\ln \frac{1}{\pi(\hat{h}_S^*)\delta}}{2n}}$

Application example: binary decision trees







…

$|\mathcal{H}_0| = 2$

$|\mathcal{H}_1| = 4$

$|\mathcal{H}_2| = 16$

$|\mathcal{H}_d| = 2^{|\mathcal{X}_d|} = 2^{2^d}$

An alternative representation

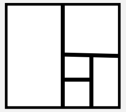
	$\mathcal{X}_2$		$\mathcal{Y}$
h_1	0	0	0
	0	1	0
	1	0	0
	1	1	0

	$\mathcal{X}_2$		$\mathcal{Y}$
h_2	0	0	1
	0	1	0
	1	0	0
	1	1	0

⋮

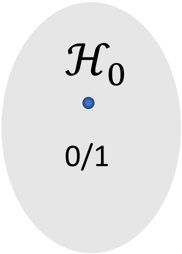
	$\mathcal{X}_2$		$\mathcal{Y}$
$h_{\vert\mathcal{H}_2\vert}$	0	0	1
	0	1	1
	1	0	1
	1	1	1

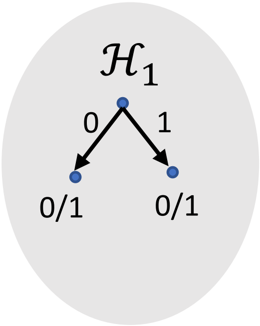
$$\sum_{d=1}^{\infty} \frac{1}{2^d} = 1$$

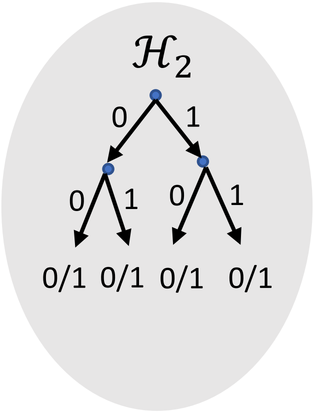


Application example: binary decision trees

$\pi(\mathcal{H}_0) = \frac{1}{2}$ $\pi(\mathcal{H}_1) = \frac{1}{4}$ $\pi(\mathcal{H}_2) = \frac{1}{8}$ $\pi(\mathcal{H}_d) = \frac{1}{2^{d+1}}$







…

$|\mathcal{H}_0| = 2$ $|\mathcal{H}_1| = 4$ $|\mathcal{H}_2| = 16$ $|\mathcal{H}_d| = 2^{2^d}$

Occam: pick $\pi(h)$, such that $\sum_{h \in \mathcal{H}} \pi(h) \leq 1$. With probability at least $1-\delta$, for all $h \in \mathcal{H}$

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}$$

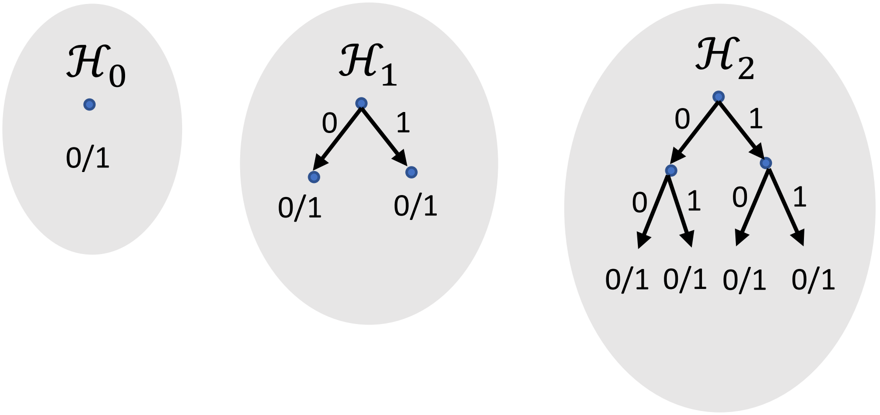
- $\mathcal{H} = \cup_{d=0}^{\infty} \mathcal{H}_d$
- $d(h)$ - depth of tree h
- $\pi(h) = \pi(\mathcal{H}_{d(h)}) \frac{1}{|\mathcal{H}_{d(h)}|}$

  $= \frac{1}{2^{d(h)+1}} \frac{1}{2^{2^{d(h)}}}$
- $\sum_{h \in \mathcal{H}} \pi(h) = \sum_{d=0}^{\infty} \sum_{h \in \mathcal{H}_d} \pi(h) = 1$

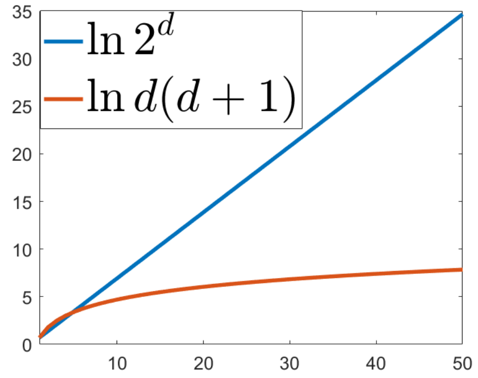
With probability at least $1-\delta$, for all $h \in \mathcal{H}$:

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{(2^{d(h)} + d(h) + 1)\ln(2) + \ln\frac{1}{\delta}}{2n}}$$

The choice of $\pi(h)$



- $\pi(h) = \pi\left(\mathcal{H}_{d(h)}\right) \frac{1}{\left|\mathcal{H}_{d(h)}\right|}$
- $\frac{1}{\left|\mathcal{H}_{d(h)}\right|}$
  - Permutation-symmetric trees get the same prior
  - In absence of prior knowledge, no reason to discriminate (structurally symmetric prior)
  - If we had some prior knowledge, we could incorporate it into the prior

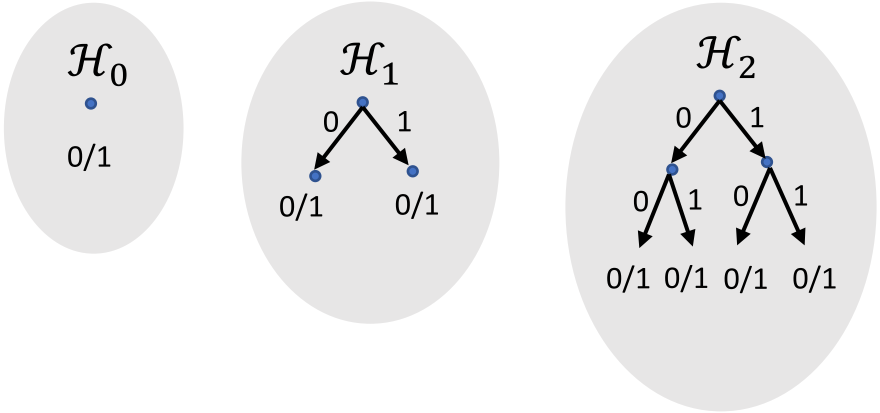


- $\pi(\mathcal{H}_d)$
  - Any series that sum up to 1 are acceptable
  - Alternative series: $\sum_{d=1}^{\infty} \frac{1}{d(d+1)} = \sum_{d=1}^{\infty} \left(\frac{1}{d} - \frac{1}{d+1}\right) = 1$
  - The bound with $\pi(\mathcal{H}_d) = \frac{1}{2^{d+1}}$: $\quad L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\left(2^{d(h)} + d(h) + 1\right)\ln(2) + \ln\frac{1}{\delta}}{2n}}$
  - The bound with $\pi(\mathcal{H}_d) = \frac{1}{(d+1)(d+2)}$: $\quad L(h) \leq \hat{L}(h,S) + \sqrt{\frac{2^{d(h)}\ln(2) + \ln\left((d(h)+1)(d(h)+2)\right) + \ln\frac{1}{\delta}}{2n}}$
  - Here $|\mathcal{H}_d| = 2^{2^d}$ is the dominant term, but elsewhere the choice of a series can make a big difference
  - $\sum_{d=1}^{\infty} \frac{1}{d(d+1)}$ is almost as close to uniform $\frac{1}{d}$ as it may get: $\ln d(d+1) \approx 2\ln d$. Uniform makes no prior assumptions

Occam and estimation-approximation trade-off

- Occam's razor resolves the estimation-approximation trade-off
  - We do not need to select $|\mathcal{H}|$ before learning starts, we select data-dependently

- This is achieved by h-dependent balancing of precision $\sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}$ and confidence $\pi(h)\delta$



$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta$$

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta$$

The "prior" $\pi(h)$

- The "complexity" $\pi(h)$ is defined for each h individually, before learning starts
- Large $\pi(h)$ gives a small complexity term, but due to $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$ it cannot be large for too many h
- The bound is only meaningful for h with $\pi(h) \gg e^{-n}$
  - So effectively we work with at most e^n prediction rules
- If $\pi(h)$ is large for h with low $\hat{L}(h,S)$, we obtain a good bound
- But if $\pi(h)$ is small for all h with low $\hat{L}(h,S)$, then the bound is loose
- Therefore, we do not want $\pi(h)$ to be concentrated on too few h
- The bound may be good or bad, but it is always valid (unlike Bayesian approaches)
- $\pi(h)$ is an auxiliary construction in derivation of the bound
  - It can be used to encode prior knowledge, but it is not a "belief" in the Bayesian sense

Mid-Summary

- Occam's razor: $\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta$

- Automatically addresses the estimation-approximation trade-off
  - Adaptive data-dependent selection guided by $\pi(h)$

- Example - binary decision trees:

$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{2^{d(h)}\ln(2) + \ln\big((d(h)+1)(d(h)+2)\big) + \ln\frac{1}{\delta}}{2n}}\right) \leq \delta$$

  - Data-dependent selection of depth

Outline

- Supervised learning – general background on generalization guarantees
- Occam's razor – "the little brother of PAC-Bayes" [skipped]
- **PAC-Bayesian analysis (including distinctions with Bayesian learning)**
- Recursive PAC-Bayes – sequential prior updates
- Weighted majority votes (if we reach it…)

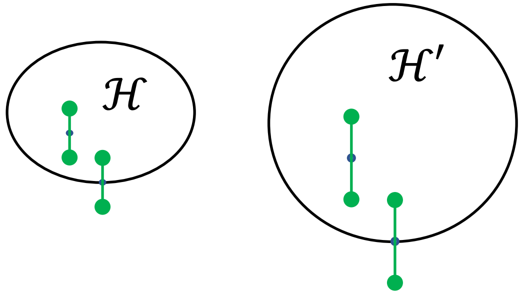
From Occam to PAC-Bayes

- Occam can only handle countable selection
  - because it is based on a union bound

- PAC-Bayes handles uncountable selection
  - by using change-of-measure inequality instead of the union bound

- and provides refined measure of selection

PAC-Bayesian Analysis



- Selection → increases estimation error

- PAC-Bayesian analysis
  - Randomized classifiers → active avoidance of selection → reduces estimation error
  - The idea: instead of committing to a particular classifier, return a distribution over classifiers (avoid commitment)
  - For example: if two classifiers have the same empirical error, do not select among them, but return a 50/50 distribution
    - Stays at the same level of approximation error, but reduces the estimation error
  - Can be applied to uncountably infinite $\mathcal{H}$

Randomized Classifiers

- Let ρ be a distribution on $\mathcal{H}$

- Randomized classification:
  1. Sample $h \sim \rho(h)$
  2. Observe X
  3. Return $h(X)$

- ρ is a *randomized classifier* / *Gibbs classifier*

- Expected error: $\mathbb{E}_{h\sim\rho(h)}[L(h)]$
- Empirical error: $\mathbb{E}_{h\sim\rho(h)}\left[\hat{L}(h,S)\right]$

Interpretation-friendly PAC-Bayes bound

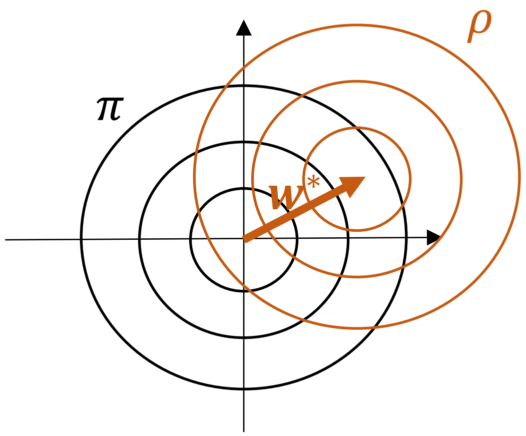
- With probability at least $1-\delta$, for all ρ:

$$\mathbb{E}_\rho[L(h)] \leq \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \sqrt{\frac{2\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}} + \frac{2\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}$$

- Pick ρ that optimizes the trade-off between $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ and $\mathrm{KL}(\rho||\pi)$
  - $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ - assign high weight to h with small $\hat{L}(h,S)$
  - $\mathrm{KL}(\rho||\pi)$ - stay close to π in the KL sense
  - Extreme case: if $\rho = \pi$, then $\mathrm{KL}(\rho||\pi) = 0$. No selection, no penalty!
  - Fast rates: small $\hat{L}(h,S)$ allows more aggressive deviation from π

Working with the bound

$$\mathbb{E}_\rho[L(h)] \le \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \sqrt{\frac{2\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}} + \frac{2\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}$$



- Select π. Example: $\pi(h_w) = \mathcal{N}(0, I)$
- Select ρ. Example: $\rho(h_w) = \mathcal{N}(w^*, I)$
- Calculate $\mathrm{KL}(\rho||\pi)$. In the example: $\mathrm{KL}(\rho||\pi) = ||w^*||^2$
- Calculate $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$
  - The challenging part

PAC-Bayes-kl inequality

- Binary kl divergence: $p, q \in [0,1]$ biases of Bernoulli distribution
  - $\mathrm{kl}(p||q) = \mathrm{KL}\big((1-p,p)||(1-q,q)\big) = (1-p)\ln\frac{1-p}{1-q} + p\ln\frac{p}{q}$

- Theorem: For any "prior" distribution π on $\mathcal{H}$ that is independent of S

$$\mathbb{P}\left(\exists\rho: \mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \geq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$$

- The interpretation-friendly bound shown earlier follows from PAC-Bayes-kl by refined Pinsker's inequality: if $\mathrm{kl}(p||q) \leq \varepsilon$, then $q \leq p + \sqrt{2p\varepsilon} + 2\varepsilon$

- $\mathbb{P}\left(\exists\rho:\ \mathbb{E}_\rho[L(h)] \geq \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \sqrt{\frac{2\mathbb{E}_\rho[\hat{L}(h,S)]\left(\mathrm{KL}(\rho||\pi)+\ln\frac{2\sqrt{n}}{\delta}\right)}{n}} + \frac{2\left(\mathrm{KL}(\rho||\pi)+\ln\frac{2\sqrt{n}}{\delta}\right)}{n}\right) \leq \delta$

Key tool for PAC-Bayes proofs: change of measure

- Donsker-Varadhan's variational formula:
  - $\ln \mathbb{E}_{X\sim\pi}[e^X] = \sup_{\rho \ll \pi}\left(\mathbb{E}_{X\sim\rho}[X] - \mathrm{KL}(\rho||\pi)\right)$

- Change of measure inequality:

  For any f, ρ, and π: $\mathbb{E}_{h\sim\rho}[f(h)] - \mathrm{KL}(\rho||\pi) \leq \ln \mathbb{E}_{h\sim\pi}\left[e^{f(h)}\right]$

Change of measure: $\mathbb{E}_\rho[f(h,S)] - \mathrm{KL}(\rho||\pi) \le \ln \mathbb{E}_\pi\left[e^{f(h,S)}\right]$

Markov:

$$\mathbb{P}\left(X \ge \frac{\mathbb{E}[X]}{\delta}\right) \le \delta$$

PAC-Bayes Lemma

- PAC-Bayes Lemma: for π independent of S
  - $\mathbb{P}\left(\exists \rho : \mathbb{E}_\rho[f(h,S)] \ge \mathrm{KL}(\rho||\pi) + \ln \frac{\mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}\right) \le \delta$
- Proof

$$\mathbb{P}\left(\exists \rho : \mathbb{E}_\rho[f(h,S)] \ge \mathrm{KL}(\rho||\pi) + \ln \frac{\mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}\right)$$

$$= \mathbb{P}\left(\exists \rho : \mathbb{E}_\rho[f(h,S)] - \mathrm{KL}(\rho||\pi) \ge \ln \frac{\mathbb{E}_S\left[\mathbb{E}_\pi\left[e^{f(h,S)}\right]\right]}{\delta}\right) \quad \text{(independence of } \pi \text{ and } S\text{)}$$

$$\le \mathbb{P}\left(\mathbb{E}_\pi\left[e^{f(h,S)}\right] \ge \frac{\mathbb{E}_S\left[\mathbb{E}_\pi\left[e^{f(h,S)}\right]\right]}{\delta}\right) \quad \text{(change of measure)}$$

$$\le \delta \quad \text{(Markov's inequality)}$$

- Change of measure *deterministically* relates $\mathbb{E}_\rho[f(h,S)] - \mathrm{KL}(\rho||\pi)$ for all ρ to a single quantity $\ln \mathbb{E}_\pi\left[e^{f(h,S)}\right]$. Probabilistic argument is applied only to $\mathbb{E}_\pi\left[e^{f(h,S)}\right]$.
- Change of measure serves as a continuous replacement to the union bound.

Proof of PAC-Bayes-kl

PAC-Bayes-kl:

$$\mathbb{P}\left(\exists\rho: \mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \geq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$$

PAC-Bayes Lemma:

$$\mathbb{P}\left(\exists\rho: \mathbb{E}_\rho[f(h,S)] \geq \mathrm{KL}(\rho||\pi) + \ln\frac{\mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}\right) \leq \delta$$

- Take $f(h,S) = n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)$
- kl-Lemma: $\mathbb{E}_S\left[e^{n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)}\right] \leq 2\sqrt{n}$
- Proof of PAC-Bayes-kl:

$$\mathbb{P}\left(\exists\rho: \mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \geq \frac{\mathrm{KL}(\rho||\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n}\right)$$

$$\leq \mathbb{P}\left(\exists\rho: \mathbb{E}_\rho\left[n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)\right] \geq \mathrm{KL}(\rho||\pi) + \ln\frac{\mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)}\right]\right]}{\delta}\right)$$ (convexity of kl + kl-Lemma)

$$\leq \delta$$ (PAC-Bayes Lemma)

Modularity of PAC-Bayes

$$\text{PAC-Bayes Lemma: } \mathbb{P}\left(\exists\rho: \mathbb{E}_\rho[f(h,S)] \geq \mathrm{KL}(\rho||\pi) + \ln\frac{\mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}\right) \leq \delta$$

- Different choices of $f(h,S)$ give divergence measures between $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ and $\mathbb{E}_\rho[L(h)]$
  - PAC-Bayes-kl: $f(h,S) = n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)$
    - kl-Lemma: $\mathbb{E}_S\left[e^{n\mathrm{kl}\left(\hat{L}(h,S)||L(h)\right)}\right] \leq 2\sqrt{n}$
  - PAC-Bayes-Hoeffding: $f(h,S) = n\lambda\left(L(h) - \hat{L}(h,S)\right)$
    - Hoeffding's Lemma: $\mathbb{E}_S\left[e^{n\lambda\left(L(h)-\hat{L}(h,S)\right)}\right] \leq e^{\frac{n\lambda^2}{8}}$
  - PAC-Bayes-Bernstein, PAC-Bayes-Bennett, PAC-Bayes-Unexpected-Bernstein, …

- Different choices of ρ and π give different regularizations.
  - Gaussian prior and posterior → regularization by $||w||^2 = \sum_{i=1}^d w_i^2$
  - Laplacian prior and posterior → regularization by $||w||_1 = \sum_{i=1}^d |w_i|$

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] || \mathbb{E}_\rho[L(h)]\right) \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}$$

Minimization of the bound

- Relaxation: PAC-Bayes-λ (based on refined Pinsker's inequality)

$$\mathbb{E}_\rho[L(h)] \leq \frac{\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}$$

- For a fixed ρ convex in λ, for a fixed λ convex in ρ
- Apply alternating minimization
  - $\rho_\lambda^*(h) = \frac{\pi(h)e^{-n\lambda\hat{L}(h,S)}}{\mathbb{E}_\pi\left[e^{-n\lambda\hat{L}(h,S)}\right]}$ (Gibbs distribution)
  - Holds for any $\mathcal{H}$, but computationally tractable only for finite $\mathcal{H}$
  - $\lambda_\rho^* = \frac{2}{\sqrt{\frac{2n\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{\mathrm{KL}(\rho||\pi)}+1}+1} \in (0,1]$
- The bound is *not* jointly convex in ρ and λ
  - Convergence to a *local* minimum, but in many practical cases still global

$$\mathbb{P}\left(\exists\rho\colon \mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \geq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$$

PAC-Bayes vs. Bayesian learning

- PAC-Bayesian bounds
  - $\pi(h)$ is an auxiliary construction in the proof; the bounds always hold
  - High-probability guarantee on the distance between $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ and $\mathbb{E}_\rho[L(h)]$
  - The loss function $\ell(h(X),Y)$ is a central element and impacts the bound minimizer ρ^*
  - Holds for all $\rho(h)$
  - Typically assumes $\ell(h(X),Y) \in [0,1]$. Otherwise requires smoothing of ρ or assumptions ensuring concentration of $\hat{L}(h,S)$ around $L(h)$
  - $\rho^*(h) \propto \pi(h)e^{-n\lambda\hat{L}(h,S)}$ and typically $\lambda < 1$
  - If ℓ is not the negative log likelihood, then the shape of $\rho^*(h)$ may be altogether different
  - A bound on $\mathrm{KL}(\rho||\pi)$ is sufficient, no need in explicit $\pi(h)$. E.g., distribution-dependent π.

- Bayesian learning
  - $\pi(h)$ is a prior “belief”.
  - The Bayes rule provides a way to update prior “belief” to a posterior “belief” given evidence (data S)

$$\rho(h|S) = \frac{\pi(h)\mathbb{P}(S|h)}{\mathbb{P}(S)}$$

  - The loss function is not part of the basic formulation
  - The posterior is a conditional distribution $\rho(h|S)$
  - Not directly concerned with concentration, so unboundedness of ℓ is not an issue
  - If ℓ is negative log-likelihood, then $\mathbb{P}(S|h) = e^{-n\hat{L}(h,S)}$ and $\rho(h|S) \propto \pi(h)e^{-n\hat{L}(h,S)}$
  - Requires explicit $\pi(h)$ to update $\rho(h|S)$

Mid-summary

- PAC-Bayes-kl bound: with probability at least $1-\delta$, for all ρ

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]||\mathbb{E}_\rho[L(h)]\right) \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n}$$

- Interpretation-friendly relaxation

$$\mathbb{E}_\rho[L(h)] \leq \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \sqrt{\frac{2\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}} + \frac{2\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}$$

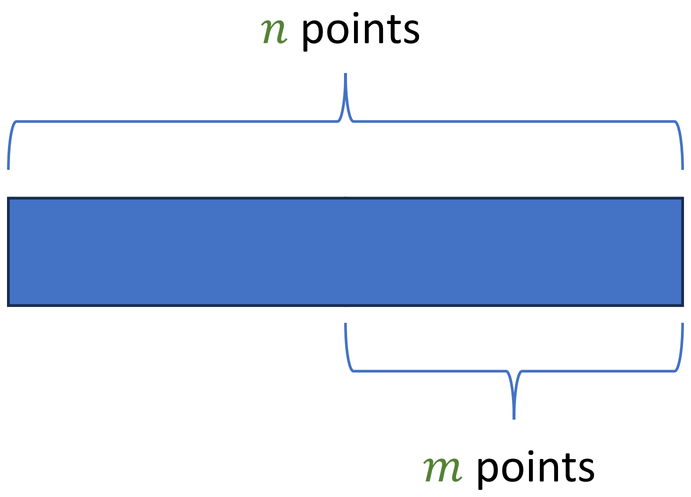
- Optimization-friendly relaxation

$$\mathbb{E}_\rho[L(h)] \leq \frac{\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}$$

Outline

- Supervised learning – general background on generalization guarantees
- Occam's razor – "the little brother of PAC-Bayes" [skipped]
- PAC-Bayesian analysis (including distinctions with Bayesian learning)
- **Recursive PAC-Bayes – sequential prior updates**
- Weighted majority votes (if we reach it…)

Data-Informed Priors



- $\mathbb{P}\left(\exists\rho: \mathrm{kl}\left(\mathbb{E}_\rho[\hat{L}(h,S)]\,||\,\mathbb{E}_\rho[L(h)]\right) \geq \frac{\mathrm{KL}(\rho||\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$

- Idea: use part of the data to construct (data-informed) π and the rest of the data to compute the bound

- Advantage: "good" π

- Disadvantage: the denominator of the bound decreases from n to m

- Challenge: the confidence information on the prior is lost
  - How much data were used to construct the prior?
  - Was it constructed in a single step or multiple steps?
  - Multi-step processing is not very helpful

Recursive PAC-Bayes

- Recursive loss decomposition

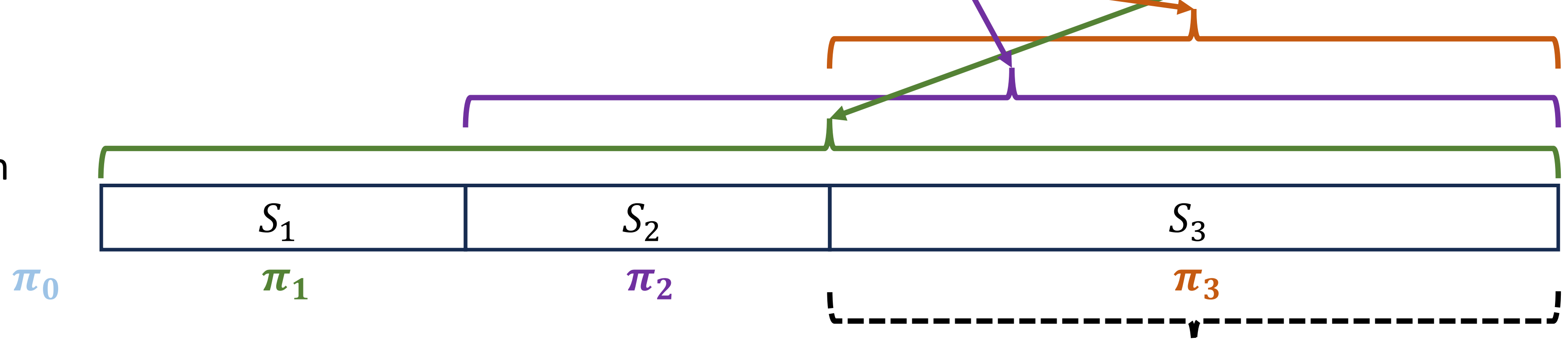
$$\mathbb{E}_{\pi_t}[L(h)] = \mathbb{E}_{\pi_t}\underbrace{\left[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]\right]}_{\substack{\text{Excess loss} \\ \text{(small variance)}}} + \gamma_t \underbrace{\mathbb{E}_{\pi_{t-1}}[L(h')]}_{\text{Decompose recursively}}$$

$$\mathbb{E}_{\pi_t}[L(h)] = \mathbb{E}_{\pi_t}\left[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]\right] + \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]$$

Illustration: 3-terms Recursive Decomposition

$$\mathbb{E}_{\pi_3}[L(h)] = \mathbb{E}_{\pi_3}\left[L(h) - \gamma_3 \mathbb{E}_{\pi_2}[L(h')]\right] + \gamma_3 \overbrace{\left(\mathbb{E}_{\pi_2}\left[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h')]\right] + \gamma_2 \mathbb{E}_{\pi_1}[L(h)]\right)}^{\mathbb{E}_{\pi_2}[L(h)]}$$

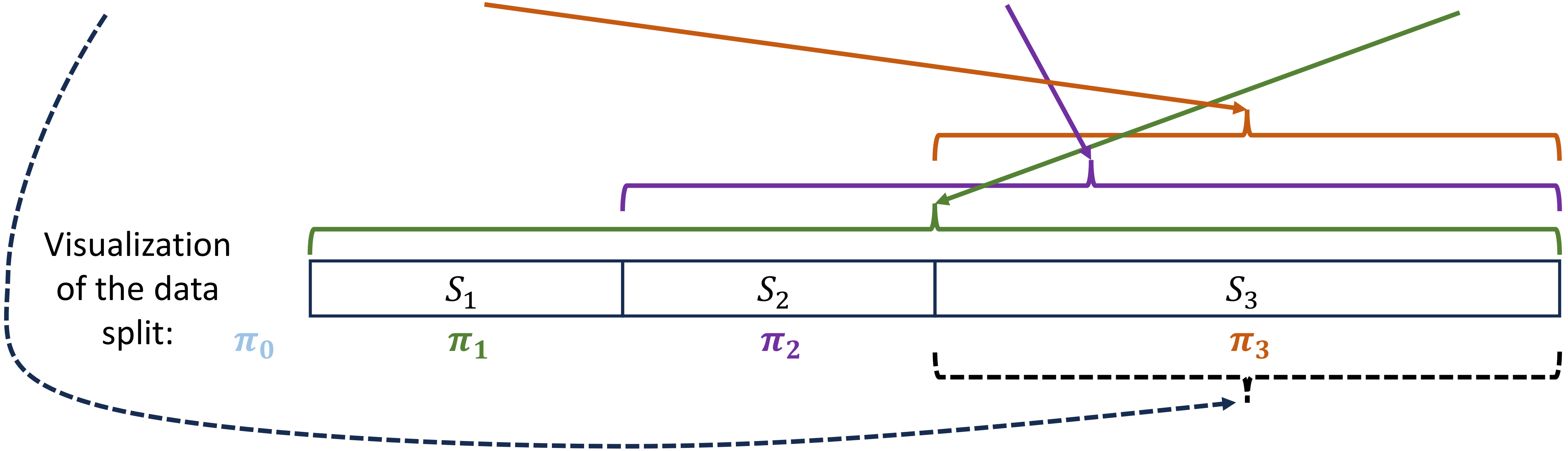
Visualization of the data split:



$$\mathbb{E}_{\pi_t}[L(h)] = \mathbb{E}_{\pi_t}\left[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]\right] + \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]$$

Illustration: 3-terms Recursive Decomposition

$$\mathbb{E}_{\pi_3}[L(h)] = \mathbb{E}_{\pi_3}\left[L(h) - \gamma_3 \mathbb{E}_{\pi_2}[L(h')]\right] + \gamma_3 \overbrace{\left(\mathbb{E}_{\pi_2}\left[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h')]\right] + \gamma_2 \mathbb{E}_{\pi_1}[L(h)]\right)}^{\mathbb{E}_{\pi_2}[L(h)]}$$

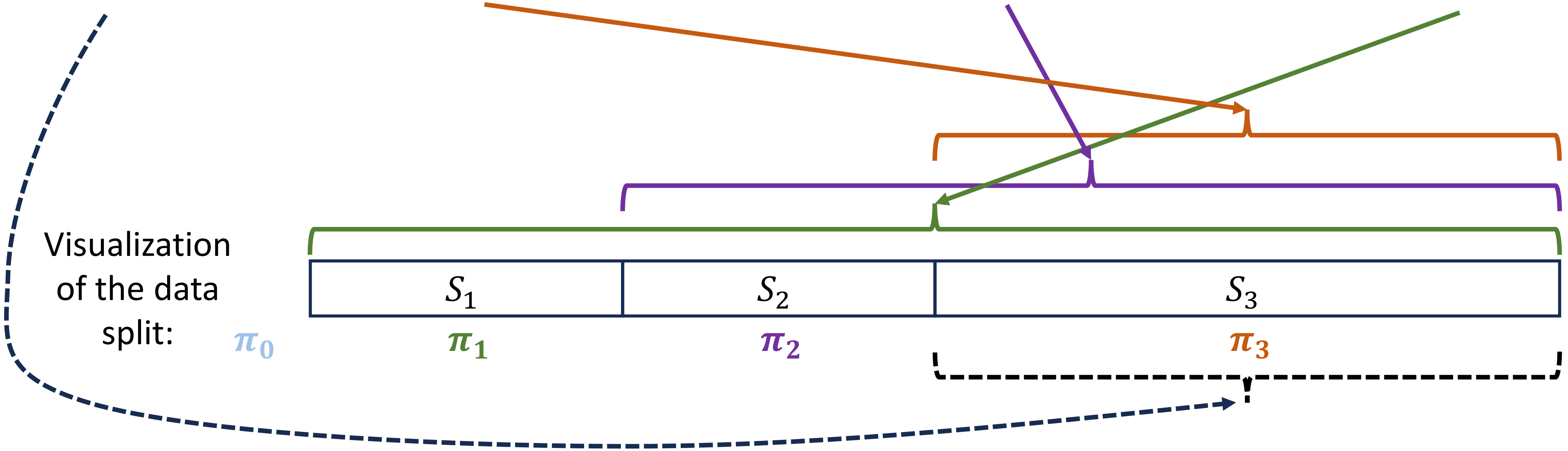


- Early steps:
  - Poor $\pi_{t-1} \Rightarrow$ large $\mathrm{KL}(\pi_t || \pi_{t-1})$
  - Large denominator
  - Small $\gamma_t \cdot \ldots \cdot \gamma_T$
- Late steps:
  - Good $\pi_{t-1} \Rightarrow$ small $\mathrm{KL}(\pi_t || \pi_{t-1})$
  - Benefit from small variance of the excess loss

$$\mathbb{E}_{\pi_t}[L(h)] = \mathbb{E}_{\pi_t}\left[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]\right] + \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]$$

Illustration: 3-terms Recursive Decomposition

$$\mathbb{E}_{\pi_3}[L(h)] = \mathbb{E}_{\pi_3}\left[L(h) - \gamma_3 \mathbb{E}_{\pi_2}[L(h')]\right] + \gamma_3 \overbrace{\left(\mathbb{E}_{\pi_2}\left[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h')]\right] + \gamma_2 \mathbb{E}_{\pi_1}[L(h)]\right)}^{\mathbb{E}_{\pi_2}[L(h)]}$$



- A good way to split the data – geometric split ($|S_t| = 2|S_{t-1}|$)
  - Use little data to bring π_t to a "good region"
  - Leave sufficient data to bound the last term $\left(|S_T| = \frac{1}{2}|S|\right)$

Empirical evaluation

	MNIST			Fashion MNIST		
	Train 0-1	Test 0-1	Bound	Train 0-1	Test 0-1	Bound
Uninf.	.343 (2e-3)	.335 (3e-3)	.457 (2e-3)	.382 (2e-3)	.384 (2e-3)	.464 (2e-3)
Inf.	.377 (8e-4)	.371 (6e-3)	.408 (9e-4)	.412 (1e-3)	.413 (6e-3)	.440 (1e-3)
Inf. + Ex.	.157 (2e-3)	.151 (3e-3)	.192 (2e-3)	.280 (4e-3)	.285 (5e-3)	.342 (6e-3)
RPB $T = 2$	.143 (2e-3)	.139 (3e-3)	.321 (3e-3)	.257 (3e-3)	.266 (5e-3)	.404 (3e-3)
RPB $T = 4$	.112 (1e-3)	.109 (1e-3)	.203 (8e-4)	.203 (2e-3)	.213 (3e-3)	.293 (1e-3)
RPB $T = 6$	.103 (1e-3)	.101 (1e-3)	.166 (1e-3)	.186 (4e-4)	.198 (1e-3)	.255 (1e-3)
RPB $T = 8$	**.101 (1e-3)**	**.097 (2e-3)**	**.158 (2e-3)**	**.181 (1e-3)**	**.192 (3e-3)**	**.242 (1e-3)**

- Uninformed π
- Informed π using half of the data
- Informed π with excess loss
  - Mhammedi et al. (2019)
  - $\mathbb{E}_\rho[L(h)] = \mathbb{E}_\rho[L(h) - L(h^*)] + L(h^*)$
  - π and h^* trained on half of the data
- Recursive PAC-Bayes (RPB) with $S = S_1, \ldots, S_T$

The computational side

- For models with linear training time the overhead is small
  - Each data point is used in training just one model
- For models with superlinear training time training a sequence of small models may be cheaper than training one big model
- If finding π_t^* involves an approximation, it may be possible to trade-off computation and accuracy

Mid-summary

- Recursive loss decomposition

$$\mathbb{E}_{\pi_t}[L(h)] = \underbrace{\mathbb{E}_{\pi_t}\left[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]\right]}_{\substack{\text{Excess loss} \\ \text{(small variance)}}} + \underbrace{\gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]}_{\text{Decompose recursively}}$$

- No loss of confidence information in the prior; efficient use of the data

- Geometric splits – little data brings π_t to a good region

Outline

- Supervised learning – general background on generalization guarantees
- Occam's razor – "the little brother of PAC-Bayes" [skipped]
- PAC-Bayesian analysis (including distinctions with Bayesian learning)
- Recursive PAC-Bayes – sequential prior updates
- **Weighted majority votes (if we reach it…)**

Weighted Majority Vote

- Fundamental technique for combining predictions of multiple classifiers
- Used in Random Forests, Boosting, and other techniques
- Can be applied with heterogeneous classifiers
- Wins most ML competitions
- Can be used for derandomization of PAC-Bayes

Key power - Cancellation of errors effect

- *If* the errors are independent, they average out

Main theoretical questions

- Generalization bounds
- Optimization of weights

Weighted Majority Vote – Formal Definition

- Let ρ be a distribution on $\mathcal{H}$
- In the binary case ($\mathcal{Y} = \{\pm 1\}$)
  - $\mathrm{MV}_\rho(X) = \mathrm{sign}(\sum_h h(X)\rho(h))$
  - The label getting the higher weight
  - Ties resolved arbitrarily
- In the multiclass case (finite $\mathcal{Y}$)
  - $\mathrm{MV}_\rho(X) = \arg\max_{\mathrm{y}} \sum_{h:h(X)=y} \rho(h) = \arg\max_{\mathrm{y}} \mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) = y)]$
  - The label getting the maximal weight
  - Ties resolved arbitrarily

Bounding $L(\mathrm{MV}_\rho)$

- Basic observation:
  - If majority vote erred, then at least a ρ-weighted half of the classifiers erred
  - $\ell\big(\mathrm{MV}_\rho(X), Y\big) \leq \mathbb{1}\ \big(\underbrace{\mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)]}_{\rho-\text{weighted mass of errors}} \geq 0.5\big)$

- Basic bound:
  - $$\underbrace{L\big(MV_\rho\big)}_{\substack{expected\ loss\ of\\ \rho-weighted\\ majority\ vote}} = \mathbb{E}_{(X,Y)\sim D}\big[\ell(\mathrm{MV}_\rho(X), Y)\big]$$

$$\leq \mathbb{E}_{(X,Y)\sim D}\ [\ \mathbb{1}(\underbrace{\mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho-\text{weighted mass}\\ \text{of errors on } (X,Y)}} \geq 0.5)]$$

$$= \mathbb{P}_{(X,Y)\sim D}\ (\ \underbrace{\mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)]}_{\substack{\rho-\text{weighted mass}\\ \text{of errors on } (X,Y)}} \geq\ 0.5\ )$$

First order oracle bound ("folk theorem")

- Theorem: First order oracle bound

$$L(\mathrm{MV}_\rho) \quad \leq \quad 2 \quad \underbrace{\mathbb{E}_{h\sim\rho}[L(h)]}_{\substack{expected\ loss\ of \\ \rho-weighted \\ randomized\ classifier}}$$

- Proof:
  - The basic bound: $L(\mathrm{MV}_\rho) \leq \mathbb{P}_{(X,Y)\sim D}\left(\mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)] \geq 0.5\right)$
  - Take $Z = \underbrace{\mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)]}_{\rho-\text{weighted mass of errors on } (X,Y)}$
  - $L(\mathrm{MV}_\rho) \leq \mathbb{P}(Z \geq 0.5) \underbrace{\leq}_{\text{Markov}} 2\mathbb{E}_{(X,Y)\sim D}[Z]$

$$= 2\mathbb{E}_{(X,Y)\sim D}\left[\mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)]\right] = 2\mathbb{E}_{h\sim\rho}\left[\mathbb{E}_{(X,Y)\sim D}[\mathbb{1}(h(X) \neq Y)]\right] = 2\mathbb{E}_{h\sim\rho}[L(h)]$$

First order empirical bound

$$\underbrace{L(\mathrm{MV}_\rho) \leq 2\mathbb{E}_{h\sim\rho}[L(h)]}_{\text{First order oracle bound}}$$

$$\leq 2\left(\underbrace{\frac{\mathbb{E}_{h\sim\rho}[\hat{L}(h,S)]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}}_{\text{PAC−Bayesian bound on } \mathbb{E}_{h\sim\rho}[L(h)]}\right)$$

- Advantages:
  - Reasonably tight
- Disadvantages:
  - The oracle bound ignores correlations of errors (the main power of MV)
  - Optimization of the empirical bound leads to deterioration of the test error

Second order oracle bound

- Define *tandem loss* for pairs of classifiers h, h'
  - $\ell(h(X), h'(X), Y) = \underbrace{\mathbb{1}(h(X) \neq Y) * \mathbb{1}(h'(X) \neq Y)}_{tandem\ loss}$
  - Counts an error if both h and h' err on (X, Y)
- Expected tandem loss:
  - $L(h, h') = \mathbb{E}_{(X,Y)\sim D}[\ell(h(X), h'(X), Y)]$
- Empirical tandem loss:
  - $\hat{L}(h, h', S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), h'(X_i), Y_i)$
- Theorem: Second order oracle bound

$$L(MV_\rho) \quad \leq \quad 4\,\mathbb{E}_{(h,h')\sim\rho^2}[\ \underbrace{L(h,h')}_{\substack{expected \\ tandem\ loss}}\ ]$$

- Proof:
  - Second order Markov's inequality:
    $\mathbb{P}(Z \geq \varepsilon) \leq \mathbb{P}(Z^2 \geq \varepsilon^2) \leq \mathbb{E}[Z^2]/\varepsilon^2$
  - As before, take $Z = \mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)]$
  - $\underbrace{L(\mathrm{MV}_\rho) \leq \mathbb{P}(Z \geq 0.5)}_{\text{The basic bound}}$

$$\begin{aligned}
&\leq 4\mathbb{E}_{(X,Y)\sim D}[Z^2] \\
&= 4\mathbb{E}_{(X,Y)\sim D}\left[(\mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)])^2\right] \\
&= 4\mathbb{E}_{(X,Y)\sim D}[\mathbb{E}_{(h,h')\sim\rho^2}[\underbrace{\mathbb{1}(h(X) \neq Y) * \mathbb{1}(h'(X) \neq Y)}_{tandem\ loss}]] \\
&= 4\mathbb{E}_{(h,h')\sim\rho^2}[\mathbb{E}_{(X,Y)\sim D}[\underbrace{\mathbb{1}(h(X) \neq Y) * \mathbb{1}(h'(X) \neq Y)}_{tandem\ loss}]] \\
&= 4\mathbb{E}_{(h,h')\sim\rho^2}[L(h, h')]
\end{aligned}$$

Second order empirical bound

$$L(MV_\rho) \leq 4\mathbb{E}_{(h,h')\sim\rho^2}[L(h,h')]$$

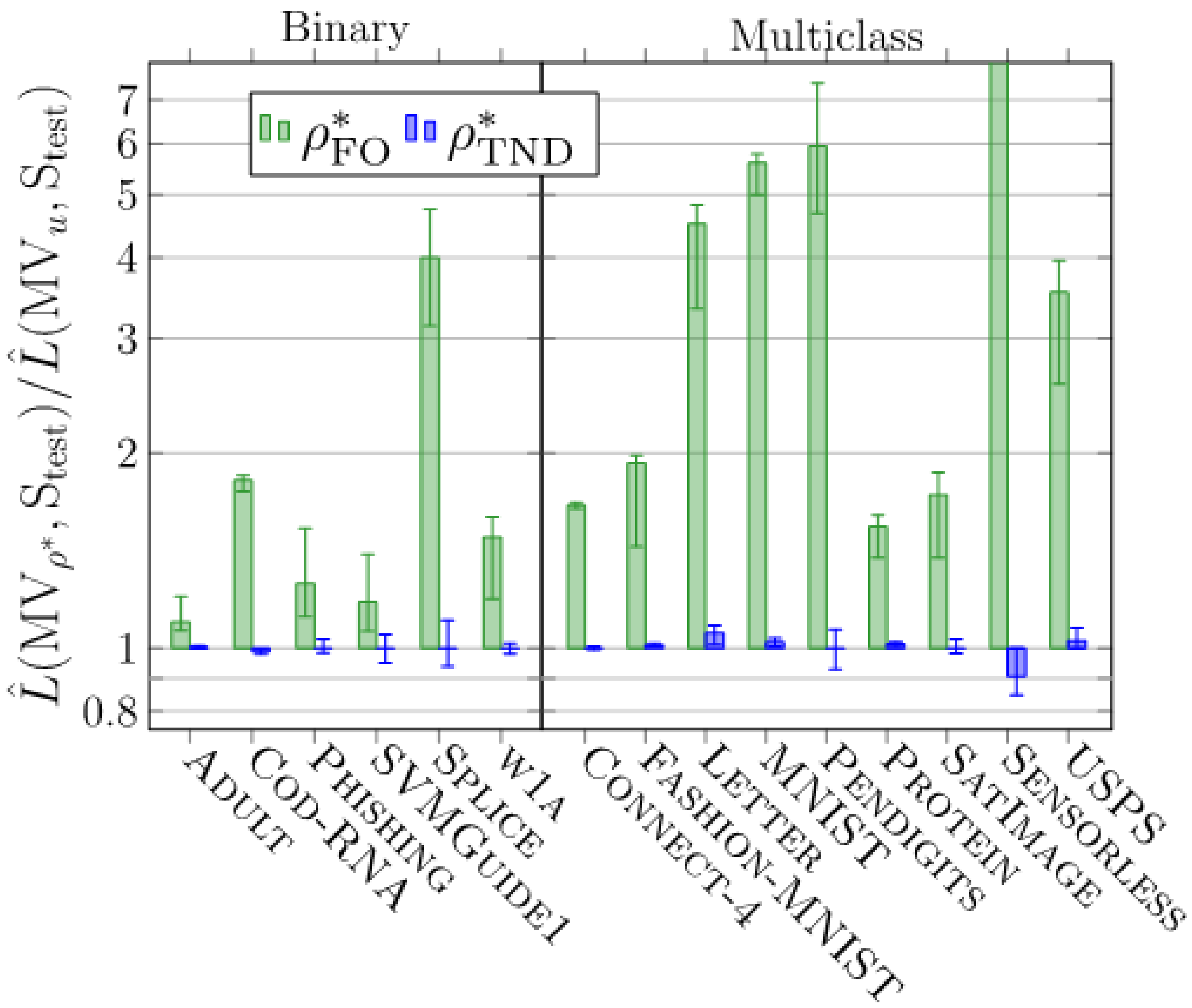
$$\leq 4\left(\underbrace{\frac{\mathbb{E}_{(h,h')\sim\rho^2}\left[\hat{L}(h,h',S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho^2\|\pi^2)+\ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}}_{\text{PAC}-\text{Bayesian bound on } \mathbb{E}_{(h,h')\sim\rho^2}[L(h,h')]}\right)$$

$$= 4\left(\frac{\mathbb{E}_{(h,h')\sim\rho^2}\left[\hat{L}(h,h',S)\right]}{1-\frac{\lambda}{2}} + \frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}\right)$$

- Advantages:
  - Takes correlation of errors into account
  - Minimization does not deteriorate the test error
- Disadvantages:
  - The tandem loss is harder to estimate (both computationally and statistically), therefore the bound is often weaker than the first order bound
    - But it is still a better bound to optimize, because it does not deteriorate the test error

Empirical evaluation

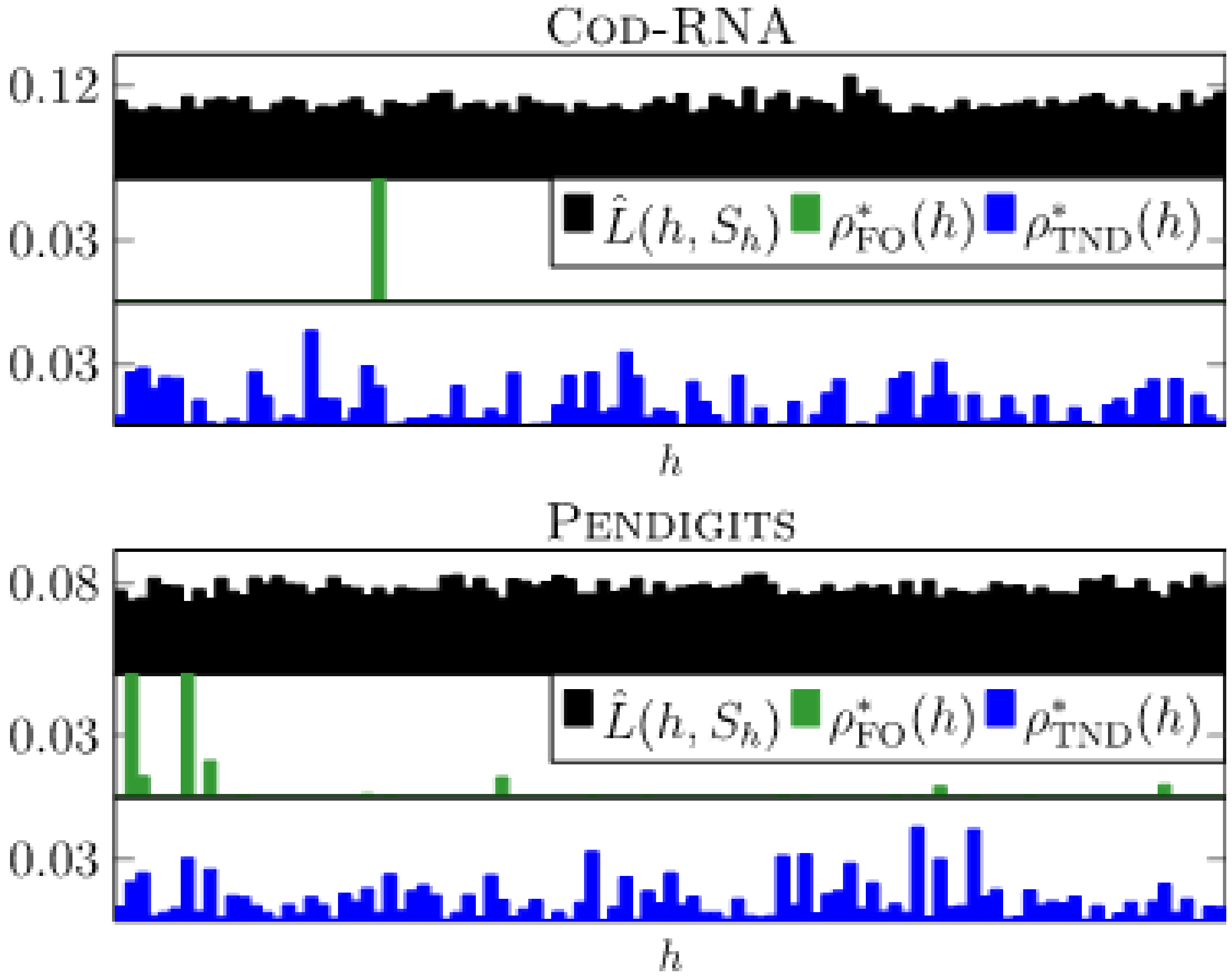
Test error of optimized majority vote over uniformly weighted majority vote baseline for the first order [FO] and the second order [TND] bound (the lower the better)



Empirical evaluation

The optimized weights ρ^* generated by the first order [FO] and the second order [TND] bound.

The result can be used for pruning the majority vote.



Mid-summary

- The basic bound:
  - $L(\mathrm{MV}_\rho) \leq \mathbb{P}_{(X,Y)\sim D}\left(\mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) \neq Y)] \geq 0.5\right)$
- First order bound
  - $\underbrace{L(\mathrm{MV}_\rho) \leq 2\mathbb{E}_{h\sim\rho}[L(h)]}_{\text{First order oracle bound}} \leq 2\left(\underbrace{\frac{\mathbb{E}_{h\sim\rho}[\hat{L}(h,S)]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}}_{\text{PAC-Bayesian bound on } \mathbb{E}_{h\sim\rho}[L(h)]}\right)$
  - Ignores correlations of errors
  - Minimization degrades the test error
- Second order bound
  - $\underbrace{L(\mathrm{MV}_\rho) \leq 4\mathbb{E}_{(h,h')\sim\rho^2}[L(h,h')]}_{\text{Second order oracle bound}} \leq 4\left(\frac{\mathbb{E}_{(h,h')\sim\rho^2}[\hat{L}(h,h',S)]}{1-\frac{\lambda}{2}} + \frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{n}}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}\right)$
  - Takes correlations of errors into account
  - Does not deteriorate the test error
  - May be weaker than the first order bound

- Follow-ups:
  - Bounds based on Chebyshev-Cantelli instead of second order Markov
    - Wu et al., NeurIPS, 2021
  - Chebyshev-Cantelli + PAC-Bayes-split-kl
    - Wu and Seldin, NeurIPS, 2022

Summary

- Occam's razor (with fast rate!) – "the little brother"

$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{2\hat{L}(h,S)\ln\frac{1}{\pi(h)\delta}}{n}} + \frac{2\ln\frac{1}{\pi(h)\delta}}{n}\right) \leq \delta$$

  - Flexible definition of complexity – prior knowledge $\pi(h)$

- PAC-Bayes

$$\mathbb{P}\left(\exists \rho:\ \mathbb{E}_\rho[L(h)] \geq \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \sqrt{\frac{2\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}} + \frac{2\left(\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{n}}{\delta}\right)}{n}\right) \leq \delta$$

  - Refined measure of selection $\mathrm{KL}(\rho||\pi)$. No selection – no penalty!

- Recursive PAC-Bayes

$$\mathbb{E}_{\pi_t}[L(h)] = \mathbb{E}_{\pi_t}\left[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]\right] + \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]$$

  - Sequential prior updates with no loss of confidence information

- Weighted majority votes

$$L\left(\mathrm{MV}_\rho\right) \leq 2\mathbb{E}_{h\sim\rho}[L(h)]$$
$$L\left(\mathrm{MV}_\rho\right) \leq 4\mathbb{E}_{(h,h')\sim\rho^2}[L(h,h')]$$

- Open question: recursive bounds for the weighted majority votes?

Further reading

- Yevgeny Seldin. Machine Learning. The science of selection under uncertainty. https://arxiv.org/pdf/2509.21547, 2025.